Available online at www.sciencedirect.com
SCIENCE DIRECT®
INTERNATIONAL JOURNAL OF APPROXIMATE REASONING
ELSEVIER International Journal of Approximate Reasoning 37 (2004) 127–144
www.elsevier.com/locate/ijar

Component-wise robust linear fuzzy clustering for collaborative filtering

Katsuhiro Honda *, Hidetomo Ichihashi

Graduate School of Engineering, Osaka Prefecture University, 1-1 Gakuen-cho, Sakai, Osaka 599-8531, Japan

Received 1 May 2003; accepted 1 February 2004
Available online 22 March 2004

Abstract

Automated collaborative filtering is a popular technique for reducing information overload and the task is to predict missing values in a data matrix. Extraction of local linear models is a useful technique for predicting the missing values. Linear models featuring local structures of the high-dimensional incomplete data set are estimated by a modified linear fuzzy clustering algorithm. Fuzzy *c*-varieties (FCV) is a linear fuzzy clustering algorithm that estimates local principal component vectors as the vectors spanning prototypes of clusters. Least squares techniques, however, often fail to account for ''outliers'', which are common in real applications. In this paper, a technique for making the FCV algorithm robust to intra-sample outliers is proposed. The objective function based on the lower rank approximation of the data matrix is minimized by a robust M-estimation algorithm that is similar to FCM-type iterative procedures. In numerical experiments, the diagnostic power of the filtering system is shown to be improved by predicting missing values using robust local linear models.

Keywords: Fuzzy *c*-varieties; Robust clustering; Collaborative filtering; Principal component analysis

* Corresponding author. Tel.: +81-72-254-9355; fax: +81-72-254-9915.
E-mail addresses: honda@ie.osakafu-u.ac.jp (K. Honda), ichi@ie.osakafu-u.ac.jp (H. Ichihashi).
URL: http://www.ie.osakafu-u.ac.jp/~honda.

0888-613X/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2004.02.001

1. Introduction

Automated collaborative filtering is a useful tool for reducing information overload and has been considerably successful in many areas [13,17,18]. Filtering systems are built on the assumption that items to be recommended are the items preferred by users who have similar interests to the active user. The task is then to predict the applicability of items to the active user based on a database of users' ratings. The problem space can be formulated as a matrix of users versus items, with each element representing the user's rating of a specific item and the goal is to predict the missing values in the data matrix [8]. GroupLens [13,17] is a well-known neighborhood-based algorithm in which the subset of neighbors is chosen considering similarities to the active user with the weighted averages of ratings used to generate predictions for the active user. However, existing methods often require a large memory space because the filtering systems have to retain all the ratings for calculating each missing value.

Honda et al. [12] proposed the use of local linear models for predicting the missing values. Linear models featuring local structures of a high-dimensional incomplete data set are estimated by a modified linear fuzzy clustering algorithm. Fuzzy c-varieties (FCV) [1,2] is a linear fuzzy clustering technique that captures the local linear structures of data sets. The prototypes of clusters are the lower dimensional linear varieties spanned by the vectors forming the orthonormal bases of principal subspaces. Because the basis vectors are derived by solving the eigenvalue problems of fuzzy scatter matrices, they are regarded as the local principal component vectors of the data sets. FCV clustering can be formulated as the extraction of local principal components based on minimization by the least squares criterion, which performs lower rank approximation of the data matrix [20]. While the objective function of the FCV algorithm is composed of the distances between data points and prototypical linear varieties, the same function can be derived from the least squares criterion that achieves component-wise approximation. In [10–12], missing values in a data matrix are ignored by multiplying "0" weights to the corresponding reconstruction errors and the local models are estimated without elimination or imputation of data elements. Once the local linear models are obtained, missing values can be predicted using a few simple linear models, and so an advantage of the method is the low memory requirements.

However, linear models based on the least squares technique are sensitive to outliers and the derived models are easily influenced by noise. In real applications, two different types of noise often exist. One occurs when the data set includes noise samples and all elements of the noise samples need to be eliminated. The other is data including intra-sample outliers. For example, in the case of information filtering, if some users give incorrect ratings to several items, the data set should be analyzed by ignoring the ratings corresponding to

noise, not by removing entire users. For the robust subspace learning of a noisy data set containing intra-sample outliers, de la Torre et al. [5,6] proposed a robust principal component analysis technique based on robust M-estimation and applied the technique to a problem in computer vision by modeling the outliers that typically occur at the pixel level.

Results of the fuzzy clustering algorithm are also influenced by outliers and several techniques for handling noise have been proposed. Davé's noise clustering [3] introduced an additional "noise cluster" so that all noise samples could be dumped into a single cluster and other clusters could capture the local structures ignoring noise samples. The possibilistic clustering technique proposed by Krishnapuram and Keller [14] tried to make data partitioning robust by using a possibilistic constraint. However, in many cases, almost every sample includes a few noise elements and conventional robust clustering methods fail to derive good results because all noise samples are eliminated, even though only a few elements of the samples are noise.

A technique for making the linear clustering algorithm robust by handling intra-sample noise and predicting missing values using robust local linear models is proposed. Introducing the M-estimation technique, the lower rank approximation is performed by ignoring only noise elements. The novel clustering algorithm is based on iterative reweighted least squares (IRLS) [9] in which additional weight parameters enable a robust approximation to be obtained by using an iterative procedure without solving a non-linear system of equations. So, the proposed technique has the following advantages. Using IRLS approach, the proposed algorithm is reduced to a simple iterative procedure, which is similar to that of the linear fuzzy clustering with missing values [10,11]. The close connection contributes to applying the proposed method to missing values estimation with robust local linear models. Then, the collaborative filtering system with the linear fuzzy clustering is modified by using robust models.

The structure of this paper is as follows: in Section 2, a new robust linear clustering approach is proposed and applied to the missing value estimation problem. In Section 3, the diagnostic power of the filtering system is shown to be improved by predicting missing values using robust local linear models. The last section offers some conclusions.

2. Simultaneous application of robust principal component analysis and fuzzy clustering

2.1. Local principal component analysis using least squares criterion

Let $X = (x_{ij})$ denote an $(n \times m)$ data matrix consisting of m-dimensional observation of n samples. The matrix is often denoted as $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$ using

n-dimensional column vectors $\boldsymbol{x}_j$ composed of the elements of the ith columns of X, or $X = (\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_n)^{\mathrm{T}}$ using m-dimensional column vectors $\tilde{\boldsymbol{x}}_i$ composed of the ith row elements of X respectively. In the following, column vectors are shown in bold face and column vectors which are composed of the row elements of a matrix are indicated by tildes ($\sim$).

The goal of the simultaneous approach to PCA and fuzzy clustering is to estimate local principal component vectors that express local linear structures of a data set by partitioning the samples into several linear clusters. FCV is a clustering method that partitions a data set into C linear fuzzy clusters. The objective function of FCV consists of distances from data points to p-dimensional prototypical linear varieties spanned by linearly independent vectors $\boldsymbol{a}_{ck}$ as follows [1,2]:

$$L_{\mathrm{fcv}} = \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci}^{\theta} \left\{ (\tilde{\boldsymbol{x}}_i - \boldsymbol{b}_c)^{\mathrm{T}} (\tilde{\boldsymbol{x}}_i - \boldsymbol{b}_c) - \sum_{k=1}^{p} \boldsymbol{a}_{ck}^{\mathrm{T}} R_{ci} \boldsymbol{a}_{ck} \right\}, \tag{1}$$

$$R_{ci} = (\tilde{\boldsymbol{x}}_i - \boldsymbol{b}_c)(\tilde{\boldsymbol{x}}_i - \boldsymbol{b}_c)^{\mathrm{T}}, \tag{2}$$

where u_{ci} denotes the membership degree of the data point $\tilde{\boldsymbol{x}}_i$ to the cth cluster and T represents the transpose of the vector. $\boldsymbol{b}_c$ is the center of the cth cluster. The weighting exponent θ is added for fuzzification. The larger θ is, the fuzzier the membership assignments are.

From the necessary condition for the optimality $\partial L_{\mathrm{fcv}}/\partial \boldsymbol{a}_{ck} = \boldsymbol{0}$, the optimal $\boldsymbol{a}_{ck}$ are derived by solving the following eigenvalue problem.

$$S_{fc} \boldsymbol{a}_{ck} = \mu_{ck} \boldsymbol{a}_{ck}, \tag{3}$$

where S_{fc} is the generalized fuzzy scatter matrix,

$$S_{fc} = \sum_{i=1}^{n} u_{ci}^{\theta} R_{ci}. \tag{4}$$

Because the optimal $\boldsymbol{a}_{ck}$ are eigenvectors corresponding to the largest eigenvalues, the vectors are regarded as the fuzzy principal component vectors extracted in each cluster considering the memberships [21].

In the same way, the cluster centers and the memberships are updated from the conditions $\partial L_{\mathrm{fcv}}/\partial \boldsymbol{b}_c = \boldsymbol{0}$ and $\partial L_{\mathrm{fcv}}/\partial u_{ci} = 0$ respectively. An iterative algorithm is used to derive the clustering result.

Honda et al. [10,11] proposed to modify the objective function using least squares criterion and discussed the relationship between the FCV clustering algorithm and the local PCA technique based on the lower rank approximation of a data matrix. Introducing memberships u_{ci}, the least squares criterion for fuzzy local PCA is defined as

$$L_{\mathrm{lsc}} = \sum_{c=1}^{C} \mathrm{tr}\Big\{ (X - Y_c)^{\mathrm{T}} U_c^{\theta} (X - Y_c) \Big\}, \tag{5}$$

where $U_c = \mathrm{diag}(u_{c1}, \ldots, u_{cn})$ and $Y_c = (y_{cij})$ denotes the lower rank approximation of the data matrix X in the cth cluster,

$$Y_c = F_c A_c^{\mathrm{T}} + \mathbf{1}_n \boldsymbol{b}_c^{\mathrm{T}}, \tag{6}$$

where $F_c = (\tilde{\boldsymbol{f}}_{c1}, \ldots, \tilde{\boldsymbol{f}}_{cn})^{\mathrm{T}}$ is the $(n \times p)$ score matrix and $A_c = (\boldsymbol{a}_{c1}, \ldots, \boldsymbol{a}_{cp})$ is the $(m \times p)$ principal component matrix of the cth fuzzy cluster. $\mathbf{1}_n$ is n-dimensional vector whose elements are all 1.

With fixed memberships, the extraction of local principal components in each cluster is equivalent to the calculation of F_c, A_c and $\boldsymbol{b}_c$ such that the least squares criterion of Eq. (5) is minimized.

From the necessary condition for the optimality of the objective function $\partial L_{\mathrm{lsc}} / \partial \boldsymbol{b}_c = \mathbf{0}$, we have

$$\boldsymbol{b}_c = (\mathbf{1}_n^{\mathrm{T}} U_c^{\theta} \mathbf{1}_n)^{-1} (X^{\mathrm{T}} - A_c^{\mathrm{T}} F_c^{\mathrm{T}}) U_c^{\theta} \mathbf{1}_n. \tag{7}$$

Usually, in PCA, the principal component score is constrained to have zero mean. Considering fuzzy partitioning, the constraint is replaced with $F_c^{\mathrm{T}} U_c^{\theta} \mathbf{1}_n = \mathbf{0}$. Then, Eq. (7) is reduced to

$$\boldsymbol{b}_c = (\mathbf{1}_n^{\mathrm{T}} U_c^{\theta} \mathbf{1}_n)^{-1} X^{\mathrm{T}} U_c^{\theta} \mathbf{1}_n. \tag{8}$$

Eq. (8) is equivalent to the updating rule for the cluster center $\boldsymbol{b}_c$ in the FCV algorithm. Substituting Eqs. (6) and (8), Eq. (5) becomes

$$L_{\mathrm{lsc}} = \sum_{c=1}^{C} \big\{ \mathrm{tr}(X_c^{\mathrm{T}} U_c^{\theta} X_c) - 2\mathrm{tr}(X_c^{\mathrm{T}} U_c^{\theta} F_c A_c^{\mathrm{T}}) + \mathrm{tr}(A_c F_c^{\mathrm{T}} U_c^{\theta} F_c A_c^{\mathrm{T}}) \big\}, \tag{9}$$

where $X_c = X - \mathbf{1}_n \boldsymbol{b}_c^{\mathrm{T}}$. From $\partial L_{\mathrm{lsc}} / \partial F_c = O$,

$$F_c A_c^{\mathrm{T}} A_c = X_c A_c. \tag{10}$$

Under the condition $A_c^{\mathrm{T}} A_c = I$, which is used in the FCV algorithm, we have $F_c = X_c A_c$ and the objective function is transformed as follows:

$$L_{\mathrm{lsc}} = \sum_{c=1}^{C} \big\{ \mathrm{tr}(X_c^{\mathrm{T}} U_c^{\theta} X_c) - \mathrm{tr}(A_c^{\mathrm{T}} X_c^{\mathrm{T}} U_c^{\theta} X_c A_c) \big\} = L_{\mathrm{fcv}}. \tag{11}$$

Therefore it can be said that Eq. (5) is equivalent to the objective function of FCV and the minimization problem is solved by computing the p largest singular values of the fuzzy scatter matrix and the associated vectors.

By the way, Eq. (5) can also be expressed as

$$L_{\mathrm{lsc}} = \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci}^{\theta} \sum_{j=1}^{m} \left(x_{ij} - \sum_{k=1}^{p} f_{cik} a_{cjk} - b_{cj} \right)^2. \tag{12}$$

This formulation means that the clustering criterion is composed of the component-wise approximation of the data matrix.

2.2. Robust local principal component analysis

When we deal with a data matrix including noise elements, local models based on least squares techniques are easily distorted. To handle noise in fuzzy clustering, Davé [3] introduced an additional "noise cluster" so that all noise samples could be dumped into that single cluster and other clusters could capture the local structures ignoring noise samples. The possibilistic clustering technique proposed by Krishnapuram and Keller [14] tried to make the data partitioning robust by using a possibilistic constraint. In the possibilistic approach, the objective function is minimized with no constraints on memberships other than the requirement that memberships should be in $[0,1]$, i.e., the membership value of a sample represents the possibility of the sample belonging to the cluster. These two robust clustering approaches have close relationship and can be identified with robust estimators based on statistics. Then, a unified view of them and a general perspective were presented by Davé and Krishnapuram [4].

However, in the analysis of large scale databases with high-dimensional observations, in many cases, almost every sample includes a few noise elements and conventional robust clustering methods fail to derive good results because all noise samples are eliminated, even though only a few elements of the samples are noise. In this section, a technique for handling intra-sample outliers in local PCA is introduced based on the component-wise least squares approximation. The M-estimation technique is a useful method for estimating robust models. The goal of the technique is to derive the solution ignoring outliers that do not conform to the assumed statistical model.

For robust principal component analysis of a noisy data set including intra-sample outliers, de la Torre et al. [5,6] proposed a robust subspace learning technique based on robust M-estimation. In the robust PCA technique, the energy function to be minimized is defined as

$$L_{\mathrm{rpca}} = \sum_{i=1}^{n} \sum_{j=1}^{m} \rho\left(x_{ij} - \sum_{k=1}^{p} f_{ik} a_{jk} - b_j \right), \tag{13}$$

where $\rho(\cdot)$ is a class of robust ρ-functions [9]. In [5,6], the Geman–McClure error function [7] is used,

$$\rho(x) = \frac{x^2}{x^2 + \sigma_j^2}, \tag{14}$$

where σ_j is a scale parameter that controls the convexity of the robust function. In the optimization process, the value of σ_j is decreased by the deterministic

annealing technique. To solve the minimization problem, the use of both the iteratively reweighted least squares (IRLS) technique [9] and the gradient descent method with a local quadratic approximation was proposed. In the following, a robust local PCA technique is proposed by introducing the ρ-function into the FCV algorithm using the least squares criterion of Eq. (12). The objective function of robust FCV with entropy regularization [15] is defined as follows:

$$L_{\mathrm{rfcv}} = \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci} \sum_{j=1}^{m} \rho\left(x_{ij} - \sum_{k=1}^{p} f_{cik} a_{cjk} - b_{cj} \right) + \lambda \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci} \log u_{ci}. \tag{15}$$

The entropy term is added for fuzzification instead of the weighting exponent of the standard FCV algorithm. The larger λ is, the fuzzier the membership assignments are. The fuzzification technique is used because of several merits, e.g., "singularities" do not occur even if several sample points are on the prototypes and cluster centers are the means of $\tilde{\boldsymbol{x}}_i$ simply weighted by u_{ci}'s.

To obtain a unique solution, the objective function is minimized under the constraint that

$$F_c^{\mathrm{T}} U_c F_c = I, \quad c = 1, \ldots, C, \tag{16}$$

$$F_c^{\mathrm{T}} U_c \mathbf{1}_n = \mathbf{0}, \quad c = 1, \ldots, C, \tag{17}$$

$$\sum_{c=1}^{C} u_{ci} = 1, \quad i = 1, \ldots, n, \tag{18}$$

and $A_c^{\mathrm{T}} A_c$ is orthogonal. Eq. (18) corresponds to the probabilistic constraint for memberships. In the proposed method, Eq. (16) is used for normalization instead of $A_c^{\mathrm{T}} A_c = I$ because that is a common practice in PCA.

The optimal solution cannot be derived from eigenvalue problems because the clustering criterion is transformed by the non-linear ρ function. Thus, the solution is derived based on the IRLS technique in which the minimization problem is formulated as a weighted least squares problem with an $(n \times m)$ weight matrix $W_c = (w_{cij})$ in each cluster. w_{cij} represents the positive weight for the previous residual $e_{cij} = x_{ij} - \sum_{k=1}^{p} f_{cik} a_{cjk} - b_{cj}$. For the Geman–McClure ρ function, the weight w_{cij} is given by

$$w_{cij} = \frac{\psi(e_{cij}, \sigma_j)}{e_{cij}}, \tag{19}$$

where

$$\psi(e_{cij}, \sigma_j) = \frac{\partial \rho(e_{cij})}{\partial e_{cij}} = \frac{2 e_{cij} \sigma_j^2}{(e_{cij}^2 + \sigma_j^2)^2}, \tag{20}$$

and the objective function is modified as follows:

$$L_{\mathrm{rfcv}'} = \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\sum_{j=1}^{m} w_{cij}e_{cij}^2 + \lambda\sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci}. \tag{21}$$

Minimization of the modified objective function approximately achieves the optimization of Eq. (15) because the first derivative with fixed w_{cij} is similar to that of L_{rfcv} as follows:

$$\begin{aligned}\frac{\partial L_{\mathrm{rfcv}'}}{\partial e_{cij}} &= 2u_{ci}w_{cij}e_{cij}\\ &\cong \frac{2u_{ci}\sigma_j^2 e_{cij}}{\left(e_{cij}^2+\sigma_j^2\right)^2} = \frac{\partial L_{\mathrm{rfcv}}}{\partial e_{cij}}.\end{aligned} \tag{22}$$

If the parameter σ_j has a large value, the modified objective function gives similar results to the FCV algorithm since all the weights w_{cij} have similar values.

To derive the optimal A_c and $\boldsymbol{b}_c$, Eq. (21) is rewritten as follows:

$$\begin{aligned}L_{\mathrm{rfcv}} = &\sum_{c=1}^{C}\sum_{j=1}^{m}(\boldsymbol{x}_j - F_c\tilde{\boldsymbol{a}}_{cj} - \mathbf{1}_n b_{cj})^{\mathrm{T}}U_cW_{cj}(\boldsymbol{x}_j - F_c\tilde{\boldsymbol{a}}_{cj} - \mathbf{1}_n b_{cj})\\ &+ \lambda\sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci},\end{aligned} \tag{23}$$

where $A_c = (\tilde{\boldsymbol{a}}_{c1},\ldots,\tilde{\boldsymbol{a}}_{cm})^{\mathrm{T}}$ and $W_{cj} = \mathrm{diag}(w_{c1j},\ldots,w_{cnj})$. From $\partial L_{\mathrm{rfcv}}/\partial\tilde{\boldsymbol{a}}_{cj} = \mathbf{0}$ and $\partial L_{\mathrm{rfcv}}/\partial b_{cj} = 0$, we have

$$\tilde{\boldsymbol{a}}_{cj} = (F_c^{\mathrm{T}}U_cW_{cj}F_c)^{-1}F_c^{\mathrm{T}}U_cW_{cj}(\boldsymbol{x}_j - \mathbf{1}_n b_{cj}), \tag{24}$$

$$b_{cj} = (\mathbf{1}_n^{\mathrm{T}}U_cW_{cj}\mathbf{1}_n)^{-1}\mathbf{1}_n^{\mathrm{T}}U_cW_{cj}(\boldsymbol{x}_j - F_c\tilde{\boldsymbol{a}}_{cj}). \tag{25}$$

In the same way, we can derive the optimal F_c and u_{ci}. Eq. (21) is equivalent to

$$\begin{aligned}L_{\mathrm{rfcv}} = &\sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}(\tilde{\boldsymbol{x}}_i - A_c\tilde{\boldsymbol{f}}_{ci} - \boldsymbol{b}_c)^{\mathrm{T}}\widetilde{W}_{ci}(\tilde{\boldsymbol{x}}_i - A_c\tilde{\boldsymbol{f}}_{ci} - \boldsymbol{b}_c)\\ &+ \lambda\sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci},\end{aligned} \tag{26}$$

and $\partial L_{\mathrm{rfcv}}/\partial\tilde{\boldsymbol{f}}_{ci} = \mathbf{0}$ and $\partial L_{\mathrm{rfcv}}/\partial u_{ci} = 0$ yields

$$\tilde{\boldsymbol{f}}_{ci} = (A_c^{\mathrm{T}}\widetilde{W}_{ci}A_c)^{-1}A_c^{\mathrm{T}}\widetilde{W}_{ci}(\tilde{\boldsymbol{x}}_i - \boldsymbol{b}_c), \tag{27}$$

$$u_{ci} = \exp\left\{-(\tilde{\boldsymbol{x}}_i - A_c\tilde{\boldsymbol{f}}_{ci} - \boldsymbol{b}_c)^{\mathrm{T}}\widetilde{W}_{ci}(\tilde{\boldsymbol{x}}_i - A_c\tilde{\boldsymbol{f}}_{ci} - \boldsymbol{b}_c)/\lambda - 1\right\}, \tag{28}$$

where $\widetilde{W}_{ci} = \mathrm{diag}(w_{ci1},\ldots,w_{cim})$.

The proposed algorithm can be written as follows.

Robust Fuzzy c-Varieties (Robust FCV) Algorithm

Step 1: Initialize $U_c, A_c, \boldsymbol{b}_c, F_c$ randomly in each cluster and normalize them so that they satisfy the constraints Eqs. (16)–(18) and $A_c^{\mathrm{T}} A_c$ is orthogonal. Choose termination thresholds ϵ_1 and ϵ_2.
Step 2: Calculate the initial W_c in each cluster.
Step 3: Update A_c using Eq. (24) and transform so that each $A_c^{\mathrm{T}} A_c$ is orthogonal.
Step 4: Update F_c using Eq. (27) and normalize so that the constraints of Eqs. (16) and (17) are satisfied.
Step 5: Update $\boldsymbol{b}_c$ using Eq. (25).
Step 6: Update U_c using Eq. (28) and normalize so that Eq. (18) holds.
Step 7: If

$$\max_{c,i} |u_{ci}^{\mathrm{NEW}} - u_{ci}^{\mathrm{OLD}}| < \epsilon_1,$$

then go to Step 8. Otherwise, return to Step 3.
Step 8: Update W_c's using Eq. (19). If

$$\max_{c,i,j} |w_{cij}^{\mathrm{NEW}} - w_{cij}^{\mathrm{OLD}}| < \epsilon_2,$$

then stop. Otherwise, return to Step 3.

The orthogonal matrices in Steps 1, 3 and 4 are obtained by such a technique as Gram-Schmidt's orthogonalization.

Usually the initial value of the parameter σ_j is large and is annealed in the optimization process. The initial partitioning can then be said to be given by the FCV algorithm.

2.3. Handling missing values and application to collaborative filtering

Aside from noise or outliers, missing values are also common in real world data analysis. In such cases, a priori knowledge about observations is often available although many noise elements are not identified prior to analysis. In [10,11], missing values in a data matrix are ignored by multiplying "0" weights over the corresponding reconstruction errors. Considering binary weights d_{ij},

$$d_{ij} = \begin{cases} 1, & x_{ij} \text{ is observed,} \\ 0, & x_{ij} \text{ is missing} \end{cases} \tag{29}$$

the objective function of FCV with missing values is defined as

$$L_{\mathrm{fcvm}} = \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\sum_{j=1}^{m} d_{ij}\left(x_{ij} - \sum_{k=1}^{p} f_{cik}a_{cjk} - b_{cj}\right)^2 + \lambda\sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci}. \tag{30}$$

The novel method is also applied to incomplete data sets. To handle missing values, the weight parameter w_{cij} is redefined as follows:

$$w_{cij} = \begin{cases} \dfrac{2\sigma_j^2}{(e_{cij}^2 + \sigma_j^2)^2}, & x_{ij} \text{ is observed,} \\ 0, & x_{ij} \text{ is missing.} \end{cases} \tag{31}$$

If all the weights w_{cij} for the observed elements x_{ij} are 1, the proposed method is equivalent to the FCV algorithm with missing values [10,11].

Once local linear models are estimated, the missing values of the data matrix X can be predicted using the corresponding elements of the approximation matrix Y_c because Y_c includes no missing values [12]. The cluster of the user is determined based on the maximum membership assignment and the missing value x_{ij} is predicted from the corresponding value y_{cij}. This means that the missing values are estimated based on the assumption that data points including missing values should exist on the nearest points to the prototypical linear varieties spanned by the local principal component vectors.

Using the local linear models, the ratings of new active users can also be predicted. Memberships and principal component scores of the new active users are estimated by Eqs. (28) and (27), respectively, and y_{cij} can be predicted using the following equation.

$$y_{cij} = \sum_{k=1}^{p} f_{cik}a_{cjk} + b_{cj}. \tag{32}$$

Even when we have an enormous number of users, the missing values can be predicted by using only Eqs. (28), (27) and the local principal component matrix A_c while memory-based algorithms such as GroupLens must retain all elements of the users versus items matrix. In this sense, the proposed prediction method is an efficient technique with fewer memory requirements.

3. Numerical experiments

3.1. Analysis of artificial data sets

Numerical experiments were performed using an artificial data set. The 3-D data set composed of 24 samples forms two lines, with the goal of the analysis being to capture these lines. Applying the FCV algorithm ($\theta = 2$) to this

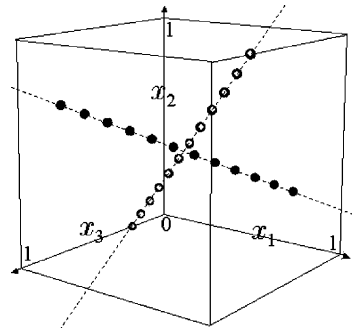

Fig. 1. Clustering result by FCV with a complete data set.

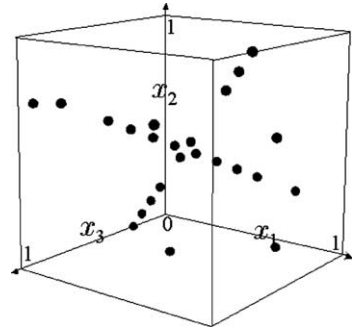

Fig. 2. 3-D plot of a noisy data set (21%).

complete data set, the samples could be partitioned into two linear clusters. Fig. 1 shows the clustering result in which ○ and ● represent the two clusters and the dotted lines are their prototypes. The prototypes represent the two lines properly.

The robustness of the proposed method was first compared to conventional noise clustering techniques. Replacing randomly selected elements with noise values, a noisy data set including 21% noise samples was constructed. Figs. 2 and 3 show 3-D plot and 2-D projections of the noisy data set. Note that each "noise sample" includes only one noise element, i.e., the noise sample points are on the line in one of three projections in Fig. 3. Fig. 4 shows clustering results derived from the standard FCV algorithm, the FCV algorithm with noise clustering mechanism and the proposed method. In Fig. 4, the dotted lines represent the prototypes of two clusters (○ and ●). $\lambda = 0.05$ and the scale parameter σ_j was annealed by the following schedule:

$$\sigma_j^2 = \frac{0.5}{\log(t+2)}, \tag{33}$$

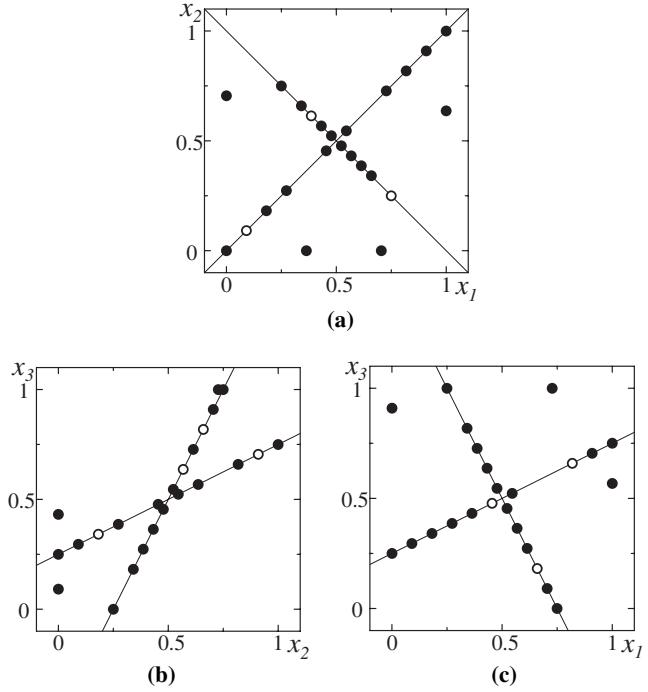

Fig. 3. 2-D projections of a noisy data set (21%): (a) $x_1 - x_2$, (b) $x_2 - x_3$ and (c) $x_1 - x_3$.

where t is the iteration index. Fig. 5 shows the trajectory of the coefficient. Although the prototypes of the standard FCV algorithm were influenced by outliers, the noise clustering version and the proposed algorithm were able to capture the two lines properly. Noise clustering ignored the samples including noise element and assigned them to the noise cluster represented by the triangles in Fig. 4. In contrast, the proposed algorithm ignored only the noise elements and assigned the noise samples into proper clusters. In this way, the algorithm provides a similar result to noise clustering except for the assignment of noise samples when the data set includes only a few noise elements.

The more difficult case in which the data set included 67% noise samples was also considered. 3-D plot of the data set is shown in Fig. 6. Because the data set has many noise samples, the local structure cannot be captured intuitively. However every noise sample is on a line in one of three projections because it includes only one noise element. The prototypes derived by FCV with noise clustering and the proposed robust FCV algorithm are shown in Fig. 7. Only the proposed method captured the two lines properly because good observations were effectively used, with only noise elements ignored. In this way, the

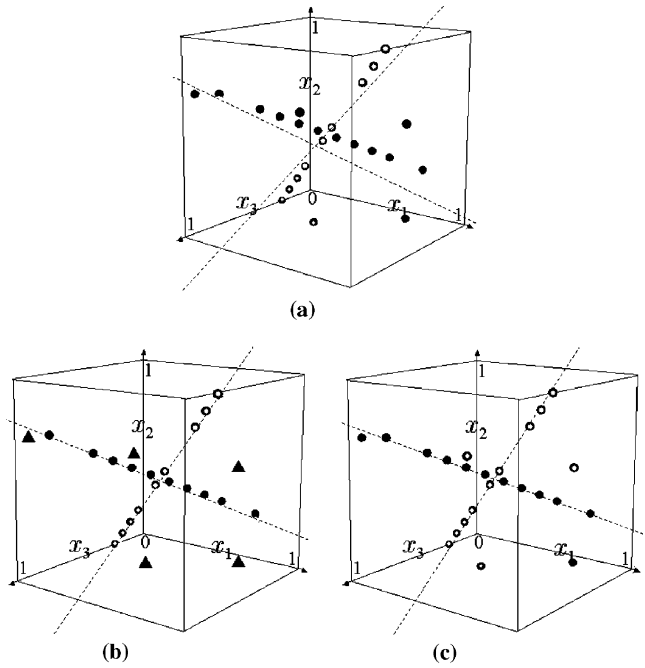

Fig. 4. Clustering results with a noisy data set (21%): (a) FCV, (b) FCV(NC) and (c) Robust FCV.

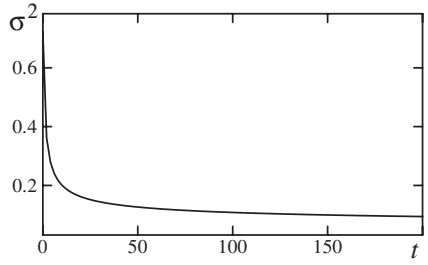

Fig. 5. Annealing schedule (Eq. (33)).

proposed method is a useful technique for analyzing data sets with intra-sample outliers.

The same experiment was then performed using an incomplete data set including both noise elements and missing values. The data set was made by withholding 10 elements from the noisy data set (21%). The ○ in Fig. 3 represents samples whose one element is missing. For example, ○ in the projection on $x_1 - x_2$ field indicates that the third element of the sample is missing. The standard FCV algorithm and the FCV with noise clustering algorithm cannot

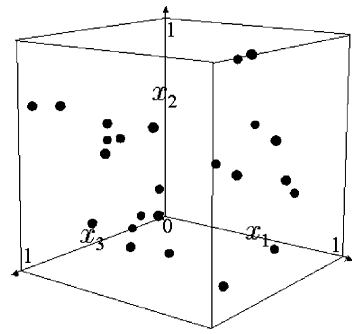

Fig. 6. 3-D plot of a noisy data set (67%).

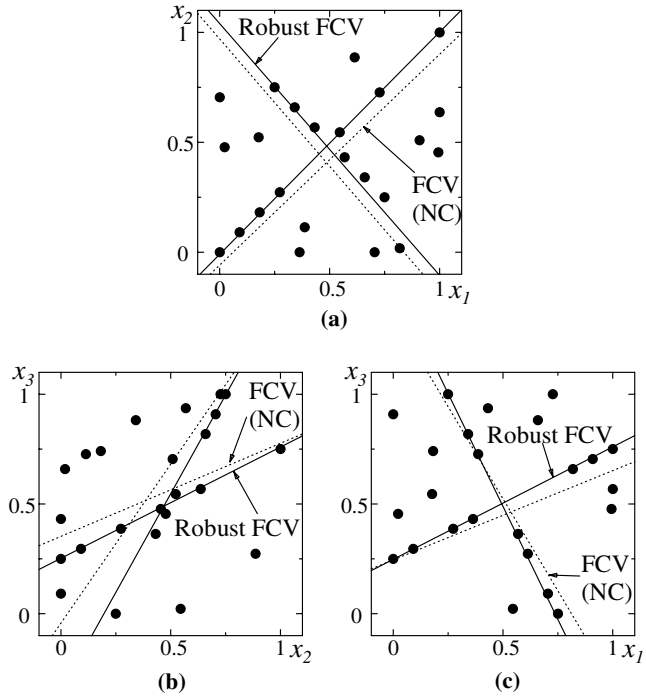

Fig. 7. Clustering results with a noisy data set (67%): (a) $x_1 - x_2$, (b) $x_2 - x_3$ and (c) $x_1 - x_3$.

capture the local structures because analysis of incomplete data is impossible without eliminating all samples that have missing values. Fig. 8 shows prototypes of the derived clusters. The ○ represents samples including one missing value. The proposed method could analyze the data set without the influence of

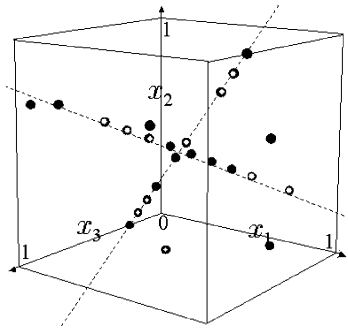

Fig. 8. Clustering result with an incomplete noisy data set (○ represent samples with one missing element).

noise or the loss of information. Because the derived prototypes were fitted properly to the (original) intrinsic structure shown in Fig. 1, they are also useful for missing value estimation, in which the missing values are assumed to be generated from the intrinsic models.

3.2. Application to collaborative filtering

The prediction technique presented in Section 2.3 was implemented for collaborative filtering and tested with ratings data collected for purposes of anonymous review from the MovieLens movie recommendation site [16]. The data set is composed of 100,000 ratings from 943 users, with every user having evaluated at least 20 ratings on a scale from 1 to 5 based on the semantic differential (SD) method. 20,000 ratings were randomly selected as test data. Only the 1240 movies that had been evaluated by at least four users were used because other movie ratings were difficult to predict from correlations among the users. In the proposed technique, users were partitioned into two clusters and one principal component vector was extracted from each cluster. Parameters were given by $\lambda = 6.0$ and the annealing schedule was

$$\sigma_j^2 = \frac{5.0}{\log(t+2)}. \tag{34}$$

In addition, ratings were also predicted using the original GroupLens [17], a non-personalized prediction method [8] and FCV with missing values [12]. The original GroupLens system predicts the jth rating of the ith active user using a weighted sum of the ratings of the other users:

$$y_{ij} = \bar{x}_i + \frac{\sum_{u=1}^{n}(x_{uj} - \bar{x}_u) \times \omega_{iu}}{\sum_{u=1}^{n} \omega_{iu}}, \tag{35}$$

Table 1
Comparison of prediction algorithms (MAE: mean absolute error, ROC: receiver operating characteristic sensitivity)

Algorithm	MAE	ROC
Non-personalized method	0.821	0.714
GroupLens	0.762	0.762
FCV with missing values	0.754	0.777
Robust FCV	0.751	0.789

where $\bar{x}_i$ is the average of the ratings voted by the ith user. The weights ω_{iu} are the similarity measures between the ith user and the uth user and the original GroupLens used Pearson correlation coefficients. In the non-personalized prediction method, ratings were computed using deviation-from-mean average over all users, i.e., missing values were calculated using Eq. (35) in which all weights were constrained to be 1. Table 1 shows a comparison of the results.

To assess the accuracy of the four prediction methods, the mean absolute error (MAE) and receiver operating characteristic (ROC) sensitivity are used. MAE is the average of the absolute deviation between prediction y_{cij} and rating x_{ij}. ROC sensitivity is a measure of the diagnostic power of a system [19]. The sensitivity refers to the probability of a randomly selected good item being accepted by the filter. The greater the value, the richer the performance. The maximum value is 1. In this experiment, movies whose ratings were larger than 3 were regarded as good items and the filtering system recommended movies whose predictions were larger than 3.5. Then, the probability was calculated as (number of recommended items whose ratings were larger than 3)/(number of items whose ratings were larger than 3). The proposed method provided the best performance, i.e., the local linear models derived by the proposed method represented the principal local features properly. In this way, the proposed method is useful for extracting robust local features of large scale databases and robust local models are available for missing value estimation. However, model parameters, such as the number of clusters C or the dimensionality of the local subspaces p must be chosen carefully. In this experiment, $(C, p) = (2, 1)$ was the best choice, but the performance severely depends on the data set. In the real application, the parameters should be determined by cross-validation techniques.

4. Conclusion

A new linear fuzzy clustering method that is robust to intra-sample noise was proposed. By introducing the "component-wise" robust M-estimation technique, the FCV algorithm was generalized to noisy data sets. Although the

objective function includes non-linear functions, additional weight parameters make it possible to derive the solution by an FCM-like iterative algorithm in which the weight parameters control the responsibility of each element in the local modeling. While conventional noise clustering techniques and possibilistic approaches ignore all "noise samples", even when the samples include only a few noise elements, the proposed method ignores only "noise elements". In this paper, the memberships were fuzzified by using entropy regularization. It can easily be shown that a similar iterative algorithm is formulated based on the standard fuzzifier used in the original FCV clustering. When a priori information about the observation is available, missing values can also be handled by constraining the associated weights to 0. Once local linear models are estimated, the missing values of the data matrix can be predicted using the corresponding elements of the approximation matrix. In numerical experiments, the diagnostic power of the filtering system was shown to be improved by predicting missing values using robust local linear models.

In spite of the usefulness, however, the FCV clustering has several drawbacks. For example, the prototypical linear varieties have infinite "size", and they may regard widely separated clusters as a single cluster. To solve the "size" problem, the convex combination of the squared distance from a data point to a cluster center and the squared distance from the same point to a linear variety was used as the clustering criterion in the Fuzzy c-elliptotypes (FCE) clustering [1,2]. However, the novel approach cannot be enhanced to the robust FCE clustering without further modification of the objective function because the clustering criterion must be written in component-wise formulation as Eq. (12). The modification is included in future works.

References

[1] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, 1981.
[2] J.C. Bezdek, C. Coray, R. Gunderson, J. Watson, Detection and characterization of cluster substructure, 2. Fuzzy c-varieties and convex combinations thereof, SIAM J. Appl. Math. 40 (2) (1981) 358–372.
[3] R.N. Davé, Characterization and detection of noise in clustering, Pattern Recogn. Lett. 12 (11) (1991) 657–664.
[4] R.N. Davé, R. Krishnapuram, Robust clustering methods: a unified view, IEEE Trans. Fuzzy Syst. 5 (1997) 270–293.
[5] F. de la Torre, M.J. Black, Robust principal component analysis for computer vision, in: Proc. of International Conference on Computer Vision, 2001, pp. 362–369.
[6] F. de la Torre, M.J. Black, A framework for robust subspace learning, Int. J. Comput. Vision 54 (2003) 117–142.
[7] S. Geman, D.E. McClure, Statistical methods for tomographic image reconstruction, Bull. Int. Statist. Inst. LII-4 (1987) 5–21.

[8] J.L. Herlocker, J.A. Konstan, A. Borchers, J. Riedl, An algorithmic framework for performing collaborative filtering, in: Proc. of Conference on Research and Development in Information Retrieval, 1999.
[9] P.W. Holland, R.E. Welsch, Robust regression using iteratively reweighted least-squares, Commun. Statist. A 6 (9) (1977) 813–827.
[10] K. Honda, H. Ichihashi, Linear fuzzy clustering techniques with missing values and their application to local principal component analysis, IEEE Trans. Fuzzy Syst. 12 (2) (2004), in press.
[11] K. Honda, N. Sugiura, H. Ichihashi, Simultaneous approach to principal component analysis and fuzzy clustering with missing values, in: Proc. of Jnt. 9th Int. Fuzzy Syst. Assoc. World Cong. and 20th North American Fuzzy Inf. Processing Soc. Int. Conf., 2001, pp. 1810–1815.
[12] K. Honda, N. Sugiura, H. Ichihashi, S. Araki, Collaborative filtering using principal component analysis and fuzzy clustering, in: Web Intelligence: Research and Development, Lecture Notes in Artificial Intelligence, vol. 2198, Springer, 2001, pp. 394–402.
[13] J.A. Konstan, B.N. Miller, D. Maltz, J.L. Herlocker, L.R. Gardon, J. Riedl, Grouplens: applying collaborative filtering to usenet news, Commun. ACM 40 (3) (1997) 77–87.
[14] R. Krishnapuram, J.M. Keller, A possibilistic approach to clustering, IEEE Trans. Fuzzy Syst. 1 (1993) 98–110.
[15] S. Miyamoto, M. Mukaidono, Fuzzy *c*-means as a regularization and maximum entropy approach, in: Proc. of the 7th International Fuzzy Systems Association World Congress, vol. 2, 1997, pp. 86–92.
[16] MovieLens Web Page, http://www.movielens.org/.
[17] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, Grouplens: an open architecture for collaborative filtering of netnews, in: Proc. of ACM Conference on Computer-Supported Cooperative Work, 1994, pp. 175–186.
[18] U. Shardanand, P. Maes, Social information filtering: algorithms for automating "word of mouth", in: Proc. of ACM Conference on Human Factors in Computing Systems, 1995, pp. 210–217.
[19] J.A. Swets, Measuring the accuracy of diagnostic systems, Science 240 (4857) (1988) 1285–1289.
[20] P. Whittle, On principal components and least square methods of factor analysis, Skand. Akt. 36 (1952) 223–239.
[21] Y. Yabuuchi, J. Watada, Fuzzy principal component analysis and its application, Biomed. Fuzzy Human Sci. 3 (1) (1997) 83–92.